IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Health Data Driven on Continuous Blood Pressure Prediction based on Gradient Boosting Decision Tree Algorithm

BING ZHANG[1,3], JIADONG REN[1,3], YONGQIANG CHENG[2], BING WANG[2] AND ZHIYAO WEI[1,3]

[1]School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei P.R.China. (e-mail: bingzhang@ysu.edu.cn)
[2]Department of Computer Science and Technology, University of Hull, Hull HU6 7RX, UK (e-mail: Y.Cheng@hull.ac.uk)
[3]The Key Laboratory for Software Engineering of Hebei Province, Qinhuangdao, Hebei 066004, China

Corresponding author: Jiadong Ren (e-mail: jdren@ysu.edu.cn).

This work was supported in part by Yorkshire Innovation Fund, UK and iMonSys ltd. UK, in part by the National Natural Science Foundation of China under Grants 61802332, 61807028, 61772451, 61772449 and 61572420, in part by the Natural Science Foundation of Hebei Province, China, under Grant F2016203330, and in part by the Doctoral Foundation Program of Yanshan University under Grant BL18012.

ABSTRACT Diseases related to issues with blood pressure are becoming a major threat to human health. With the development of telemedicine monitoring applications, a growing number of corresponding devices are being marketed, such as the usage of remote monitoring for the purposes of increasing the autonomy of the elderly and thus encouraging a healthier and longer health span. Using the machine learning algorithms to measure blood pressure at a continuous rate is a feasible way to provide models and analysis for telemedicine monitoring data and predicting blood pressure. For this paper, we applied the gradient boosting decision tree (GBDT) whilst predicting blood pressure rates based on the human physiological data collected by the EIMO device. EIMO equipment specific signal acquisition includes: ECG, PPG. In order to avoid over-fitting, the optimal parameters are selected via the cross-validation method. Consequently, our method has displayed a higher accuracy rate and better performance in calculating the mean absolute error evaluation index than methods such as: the traditional Least squares method, Ridge regression, Lasso regression, ElasticNet, SVR and KNN algorithm. When predicting the blood pressure of a single individual, calculating the systolic pressure displays an accuracy rate of above 70% and above 64% for calculating diastolic pressure with GBDT, with the prediction time being less than 0.1s. In conclusion, applying the GBDT is the best method for predicting the blood pressure of multiple individuals: with the inclusion of data such as age, body fat, ratio and height, algorithm accuracy improves, which in turn indicates that the inclusion of new features aids prediction performance.

INDEX TERMS GBDT, Pruning, Blood Pressure Prediction, Health Monitoring.

I. INTRODUCTION

DUE to the threats of incidence and mortality in the hypertensive population, cardiovascular and cerebrovascular diseases are gaining more widespread attention, especially within the context of rapid changes in the quality of diet and lifestyle amongst the general population. Non-invasive blood pressure measurement is divided into two ways: intermittent and continuous. Traditional intermittent noninvasive blood pressure measurements (such as the Riva Rocci Korotkoff method and the Oscillography method) cannot be monitored in real time, and with the presence of many influencing factors, the risk of error is bigger. On the other hand, continuous non-invasive blood pressure measurement can monitor the changes in arterial blood pressure waveform in each heart cycle, which poses an obvious advantage, whether it is performed within the context of routine home care, the monitoring of cardiovascular disease patients and even weightlessness training for astronauts. Therefore, compared with the intermittent measurement method, non-invasive blood pressure continuous measurement in clinical medicine research is becoming more vital. Traditional blood pressure measurement methods and equipment lack the ability to measure real time, continuous blood pressure. Recently, studies have shown that research is increasingly focused on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2902217, IEEE Access

personal wearable devices used for measuring, tracking and evaluating user health-related information. Common tracking devices include: Fitbit; personal accelerometers; and motion tracking devices, as well as Samsung, Apple, and smart watches, all which have relatively simple measurement functions. Such kinds of equipment are examples of technology that is advanced enough to be able to record information such as monitoring continuous activity and sleep time - this is greatly useful for both the researchers and research subjects concerned. The new smart watch for instance, includes functions such as heart rate (HR) monitoring and location tracking with built-in sensors - this thus provides users with a very convenient interface, enabling problems to be recorded at any time. Regarding the drawbacks of such devices, despite high portability and relative ease of usage, performance rates in capturing vital signs are limited - this in turn leads to a limited ability in being able to provide a complete remote monitoring program for patient management and nursing.

Gesche et al [1] predicted a blood pressure value by establishing a linear regression model for PWV (Pulse Wave Velocity) and PTT (Pulse Transit Time), however, it is not really suitable for predicting blood pressure due to high risk of erroneous measurements. YY Hsieh et al [2] predicted a blood pressure value by establishing a linear regression model for PTT of the human body. However, the PTT extracted by PPG (Photoplethysmography) contained more noise; using PTT is not an appropriate method of predicting blood pressure. Elamvazuthi et al. [3] used PID neural network to build a blood pressure management system - large amounts of data is required for the neural network but this can be difficult to accomplish in the actual working environment. Xiaohan Li et al. [4] used a context-based recursive model for blood pressure prediction. The data set is collected from the blood pressure monitor synchronised with a cell phone and stored in the cloud. This is then followed by the entire data archive in the cloud being used to train the blood pressure prediction model. Due to the model being based on 'all-the-time series' data, its accuracy rates in predicting blood pressure is high but it cannot be guaranteed during accumulation and regarding data security. Robabeh Abbasi et al. [5] proposed a time series prediction model based on multivariate fuzzy functions that may aid physicians in choosing a suitable treatment via providing a clinical diagnosis. However, the prediction needs to be completed over a long period of time, with the disadvantage of the individual difference being large the prediction accuracy decreasing with the long term usage Peng et al. [6] used characteristics of a heartbeat signal to establish a support vector regression model to achieve continuous and unarmed blood pressure predictions; the predictive value was however not very high in relation to the actual blood pressure value. Kurylyak et al. [7] used a PPG signal to establish a neural network model to predict blood pressure at a continuous rate. Unfortunately, accuracy rates were low, despite absolute error and relative error index being utilised to evaluate the Manuja et al. [8] made a summary of the current blood pressure measurement methods and proposed that the use of nonlinear algorithms such as nonlinear support vector machines and lifting methods improve blood pressure prediction rates compared to linear regression methods. Robert Munnoch et al. [9] used the EIMO device to extract the ECG, PPG, PTT, and HR characteristics of the human body. PWV was then calculated by PTT, with three linear regression models being established using PWV and HR to predict blood pressure. The three models were expressed as (1), (2) and (3).

$$P = A(PWV)^2 + B \tag{1}$$

$$P = A(HR) + B \tag{2}$$

$$P = A(PWV)^2 + B(HR) + C \tag{3}$$

For the prediction of Ps, the best model is (3), with the average absolute error being 6.15. For the prediction of low pressure Pd, the best model is (1), with the average absolute error being 8.36. However, according to the United States ANSI / AAMI SP10-1992 requirements in predicting the systolic and diastolic pressure error: the difference that is less than or equal to 5mm/Hg is accurate. The average absolute error for high and low blood pressure is greater than 5, which is higher than the error requirements. This indicates that the method of Robert Munnoch et al. failed to meet the required accuracy rates needed during when predicting blood pressure is taking place.

The above research shows that the proposed method is unable to predict blood pressure at an accurate and continuous rate. Therefore, in response to this problem, this paper proposes applying GBDT to analyze the human physiological data collected by EIMO experimental equipment and carry out blood pressure prediction. EIMO devices are able to continuously record PPG, and calculate HR and PTT respectively to ECG and PPG signals, making them appropriate instruments for carrying out the procedure of predicting blood pressure. Regarding prevention measures, the cross-validation method is used in the model training process to prevent the GBDT from fitting or underfitting. The GBDT is then compared with the traditional six machine learning algorithms (linear regression, ridge regression, SVR, ElasticNet, KNN (K-NearestNeighbor) and Lasso(Lasso regression) with two evaluation indexes (accuracy [10] and average absolute error). Thus, GBDT is the best predictor of the algorithm. For each person, the average absolute error of the high pressure is less than 5, and the average absolute error of the low pressure is less than 3: this meets the required accuracy levels needed during the prediction of high and low blood pressure rates. Using GBDT is also more time efficient which enables immediate forecasting.

II. GBDT MODEL

A. INTRODUCTION TO MODEL'S CHARACTERISTIC ATTRIBUTES

(1)PPG involves a non-invasive technique for detecting changes in blood volume in living tissues. This can be used to make predictions within numerous important health related parameters such as heart rate, hemoglobin and blood glucose levels [11], as well as predict the value of continuous BP [12]. The low frequency part of PPG contains information regarding: breath; BP control; and body temperature adjustment. The high frequency part on the other hand, contains information related to heart pulsation [13]. Cardiovascular detection methods based on PPG technology have achieved some success.

(2) The heart rate (HR) is the speed at which the heart beats, measured by the number of contractions of the heart per minute (beats per minute). It plays a significant role in human physiological functions, evaluating health status, and the analysis of cardiovascular disease. An increase or decrease in HR will correspondingly cause an increase or decrease in Ps and Pd.

(3) Pulse transit time (PTT) is the time taken for cardiac ejection and arterial pulse wave propagation from the aortic valve to the peripheral branch vessel [14-17]. In recent years, as one of the major characteristics for BP prediction, PTT has been widely applied [18-20] whilst measuring in non-invasive environments and real-time scenarios.

B. CONSTRUCTION OF BLOOD PRESSURE PREDICTION MODEL BASED ON GBDT

The classification regression tree is a supervised learning algorithm [21]. The feature space is recursively divided into several parts (or nodes) based on the relationship between an output and one or more input factors. For the characteristics of our problem, the feature space X can be regarded as a combination of n m-dimensional eigenvectors: n $>$ 0 represents the number of samples; m = 6 represents the six feature inputs; and the subscript i represents the i^{th} sample. In this paper, the high pressure Ps is taken as the object of discussion, and the research method of low pressure Pd is consistent with Ps. The feature space X can be viewed as a combination of n feature vectors X_i = {PPG, PTT, HR, age, height, BMI (body mass index)}, and the prediction samples can be represented as Y = {y_1, y_2, y_3, ..., y_n} ,where each y_i = Ps (y_i = Pd), and y_i = Ps (y_i = Pd) is a numberic variable. In this paper, we use the input feature combination {PPG, PTT, HR, age, height, BMI} to establish the classification tree for the high and low blood pressure rates respectively. During the building process, the feature and split node (data set) is selected according to the square error minimization criterion [22]. The leaf areas U={ U_1, U_2,..., U_h} of second tree were calculated when they reached a steady state. For each leaf node region U_i, there is the output of the second tree in each leaf node region c_i and the second classification and regression tree is generated. In accordance with the above steps, from the third tree until the last tree, is constructed in the same way as the second tree in this section.

C. PRUNING OF GBDT

In the process of constructing the gradient tree of recursion, max_depth, learning_rate and n_estimators determine the construction process of gradient tree. α= max_depth, β=learning_rate, γ=n_estimators. α represents the maximum depth of each tree, and the nodes are divided when the depth of a single tree is less than α. When the depth of the tree is equal to α, the data set that needs to be divided becomes the leaf node. If α is too small, the single tree cannot produce a residual current that fits well, thus producing an algorithm that is ill-fitting. If α is too large, this may be due to each of the tree training times being extended and their algorithms being too long. Therefore, the a value should neither be too big nor too small. β represents the contribution of each tree. There is a high chance of excessive noise being present in the training data of each tree, which can lead to the over-fitting of a single tree and thus cause the whole algorithm to be over-fitted. So in this paper, each tree is limited and is multiplied by a number(learning rate, β) less than 1, which reduces its influence and thus reduces the risk of the whole algorithm over-fitting. When β is too large (close to 1), the algorithm is prone to overfitting. When β is too small (close to 0), each tree has a small influence, which makes the model less likely to fit. Simultaneously as more trees need to be added, the process of learning the whole algorithm is slowed down and the resulting training time is too long. Set the m tree to be F_m(x), and the current integration model of all trees is F_{m-1}(x).The formula for integrating f_m(x) into the current F_{m-1}(x) to form a new integrated model F_m(x) is as follows:

$$F_m(x) = F_{m-1}(x) + \beta f_m(x) \tag{4}$$

γ represents the number of trees. If the value of the build tree is equal to γ, the model stops the building process, producing model fitting data: the more the tree, the better the quality of the results. If the value of γ is too big, the model experiences an overload of training data and may also pick up some noise, resulting in the model prediction ability declining. If the value of γ is too small, the learning ability of the model decreases too much and will result in a case of under-fitting.Therefore, the value of y should be chosen within an appropriate range. The final model f(x) is shown below:

$$F_\gamma(x) = \beta \sum_{i=1}^{\gamma} f_i(x) \tag{5}$$

D. GBDT MODEL FOR BP PREDICTION

The process of predicting blood pressure using gradient tree can be divided into three parts: the construction of single tree, the integration of trees and the prediction of blood pressure. The loss function of the negative gradient classification and regression tree is continually built on to carry out the fitting of the data. The data is divided into individual leaf area, and then

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2902217, IEEE Access

into each leaf node area to find the optimal output value.The construction of a single tree is like **algorithm 1**.

Algorithm 1 DecisionTreeRegressor

Input: D={(X_1, y_1), (X_2, y_2), ...,(X_n, y_n))}, X={X_1, X_2, X_3, ..., X_n} represents a feature property set, Y= {y_1, y_2, y_3, ..., y_n} represents the predicted attribute set
Output: CART model
1. current_depth = 1
2. **for** X_1 in X :
3. **for** T_j in x_i
4. **search** min(MSE);
5. **end for**
6. **end for**
7. **if** the current depth $< \alpha$
8. Divide D to D_1 and D_2
9. curren_depth += 1
10. **DecisionTreeRegressor**(D_1);
11. **DecisionTreeRegressor**(D_2);
12. **else**
13. convert D $\rightarrow$ leaf_Node;
14. **return** U

The first line of the algorithm sets the depth of the root node to 1. Lines 2 through 6 minimize the criteria by using the squared error minimization criteria to iterate through all possible split points and split attributes to obtain optimal split feature attributes and attribute values. Rows 7 and 8 indicate that when the depth of the tree is less than α, the data set is divided into two subtrees, D_1 and D_2, by the optimal splitting method. Line 9 indicates that the current depth of the tree is increased by one for every split of the data set. Lines 11 and 12 represent recursively splitting D_1 and D_2. Lines 12 and 13 indicate that if the depth of the current tree is α, the data set is transformed into a leaf node. Line 14 represents that it returns a leaf node area when building each leaf node (the range of feature attributes).

After the construction of the each classification and regression tree, it is added to the integration model and the residual and residual negative gradient direction is calculated, with a regression tree of classification fitting in the negative gradient direction of residual. After setting up γ trees, the output value of all the γ trees are in the leaf node area U={U_1, U_2,..., U_γ, U_{i1}, U_{i2},..., U_{iL}} (assuming that the i^{th} tree has L leaves) is added and is the output of the gradient tree **algorithm 2**.

Algorithm 2 Gradient tree Boosting

1. $f_0(x) = c_1$ = mean(Y)
2. for every sample :
3. $k_i = y_i$ - c_i
4. **for** $\alpha\epsilon$(a,b)
5. **for** $\alpha\epsilon$(c,d)
6. **for** $\alpha\epsilon$(e,f)
7. **for** m in (1,γ) :
8. calculate
9. $k_i = y_i$ - $F_{m-1}(x_i)$
10. r_i = -[$\frac{\partial L(y,F(x_i))}{\partial F(x_i)}$]$_{f_{m-1}(x)}$
11. combine every x_i and r_i as new data D
12. **DecisionTreeRegressor(D,emphα)**
13. **for** every U_{mi} in U_{m1}, U_{m2},..., U_{mL} // U_{mi} indecates the number of sample in the M_{th} tree's I_{th} leaf node
14. // c_{mi} indicates the predict BP value in U_{mi}
15. Treemodel =
16. TreeModel set $\leftarrow$ TreeModel
17. **end for**
18. **end for**
19. **end for**
20. **return** TreeModel set (TMS)

Lines 1 through 3 represents the initialization gradient tree. At this point, the predicted value of each sample is the mean of the high voltage value of all samples. Lines 7 through 10 represent a combination of characteristics, the negative gradient direction of the residual and predicted residuals by the integration model being computed after the last tree construction. The value of the negative gradient direction is used to combine the eigenspace X into a new data set D. Lines 12 through 14 indicate that a classification regression tree is established for D to find the leaf node area in the negative gradient direction. U={U_1, U_2,..., U_L}. Line 15 shows that the output of the integration model is the sum of the predicted values for each tree in the samples. Lines 4 through 6 and 17 to 20 represent the traversal of each feature combination, storing the the corresponding gradient tree model for each feature combination to the set collection and return.

E. MODEL OPTIMIZATION

In order to prevent model overfitting or underfitting, the fitting of the model to the data needs to be evaluated. The method to do so involves calculating the predictive ability of the model. In a prediction model, the value we want to predict is called the dependent variable, and the value used to predict it is called the explanatory variable or the independent variable. The generalization ability of the model can be evaluated by calculating the coefficient R^2 [23]. The coefficient of R^2 can then be used to measure the degree of interpretation of the target value. In this paper, the prediction value of the algorithm is explained to the target blood pressure value. The larger R^2, the higher the predicted value of blood pressure, and the stronger the generalization ability. The calculation method of R squared is formula (6).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2902217, IEEE Access

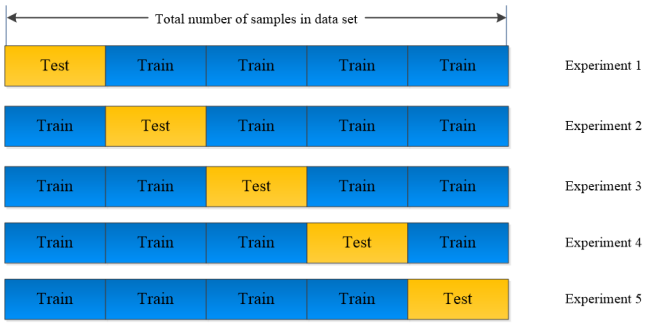


FIGURE 1: Partition diagrams of training set and verification set

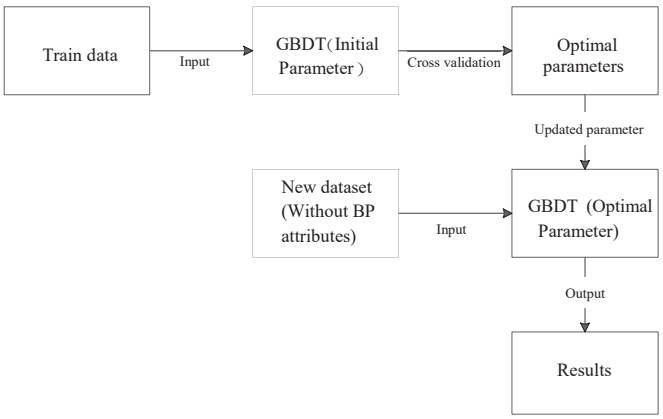


FIGURE 2: Blood pressure prediction process

$$R^2(y,m) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - m_i)^2}{\sum_{i=0}^{n-1}(y_i - k)^2} \tag{6}$$

Where y_i represents the real value of the sample i, and mi represents the predicted value of the sample i, and p represents the average value of the sample real value. Cyclic execution of the above procedure L=5 times. After computing the average value for all of R^2, this is the corresponding scores for this model, and the formula is as shown in (7).

$$scores = \frac{\sum_{i=1}^{L} R^2{}_i}{L} \tag{7}$$

The corresponding R^2 values of each parameter are calculated and the optimal parameters are selected through cross-validation. First, the training set D is divided into L large collections, then select L-1 as the training set of the model, combined with a set of D_i remaining as the validation set and test accuracy of $TreeModel_0$, **Figure 1** shows the way to divide the training set and validation set, L=5. Each model is evaluated using the formula (7) as a model evaluation method. The process is shown in **Algorithm 3**.

Algorithm 3 Finding the best parameters and model of GBDT

1. temp = 0; score = 0;
2. **for each** model in TMS :
3. **for** i in (1,L) :
4. calculate R^2
5. **end for**
6. scores = R^2 / L;
7. if $model_i$.scores $>$ temp :
8. temp = $model_i$.socre ;
9. $Treemodel_best$ = $model_i$
10. **end if**
11. **end for**
12. use F(x) to Predict Ps
13. **return** $Treemodel_{best}$'s F(x), Ps

The first line of the algorithm initializes the temp and score of the temporary variables. From the second line to the seventh line, the score value of each model is calculated. From the eighth line to the eleventh line, the model with the largest score is sought out. The twelfth line through to the thirteenth line represents the prediction of blood pressure with the optimal model, and returns the optimal model parameter combination and predicted blood pressure values. The specific stages prediction process is shown in **Figure 2**.

Figure 2 describes the whole process of blood pressure prediction, which is divided into the following steps:

(1) The processing of the data set, involves taking out the characteristic attribute data and blood pressure data from the original data file, before forming training data and test data respectively.

(2) A high pressure (or low voltage gradient) is established, which lifts the tree model to initialise (etc) the default model parameters (α =1, β =100, γ =0.1).

(3) The training set is input into the initial model and the optimal model parameters are found via cross validation. The parameters of the model are then updated as the optimal parameters. The training set is then input into the model (etc) of the optimal parameters, and the model is trained to return the trained model.

(4) The test data is input into the established model, predicting the blood pressure and outputting the rest of the prediction.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. EQUIPMENT DESCRIPTION

EIMO is among the new range of products designed to enable the user to keep track of their health and that of their family, especially within a home environment. It functions as a fully automated non-invasive device, which protects the user from any unwanted electrical discharge or painful pressure, using various sensors to take a series of measurements and displaying readings in a comprehensive, user-friendly way. It

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2902217, IEEE Access

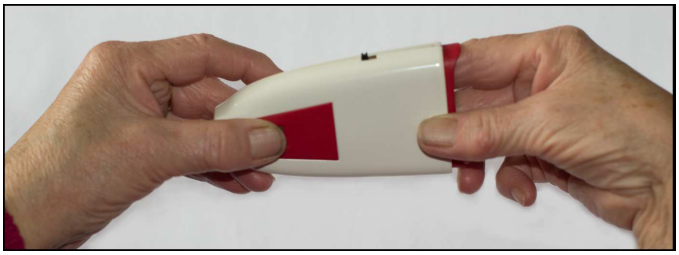


FIGURE 3: EIMO equipment

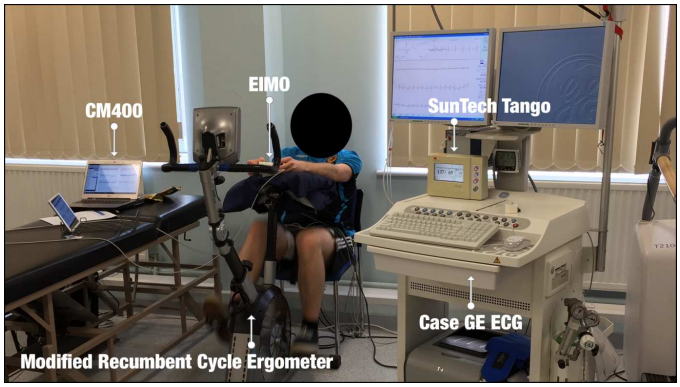


FIGURE 4: EIMO equipment collects experimental data

is classed as a medical device intended to be used to obtain readings of vital physiological signals during routine check ups and self-monitoring sessions. Therefore, it is classified as Class IIa.

The EIMO device (**Figure 3**) requires handling with both hands. The right index and middle fingers are placed into two inserts encased within the EIMO which house the electrodes for the PPG and ECG sensors. The left thumb then covers the 2nd ECG electrode at the distal end of the EIMO. There are some precautions before using EIMO equipment. Sit with EIMO on knees with the screen in a suitable position to read. Alternatively sit at a table with hands resting on the table. Do not lean on arms, EIMO is so sensitive it will pick up tiny muscle tremors, which will spoil the readings.

The battery of this device can last until approximately 7 hours. To guarantee the above operation, the BLE scheme and processing in MSP430 + CC2540 are adopted. The peak value of the signal is detected at 1KHz, and filtered as described below. These signals are then decimated to 100Hz and turned into 20 byte packets for the BLE(Bluetooth) communications. The maximum data rate that has been achieved is 3 channels running at 200 samples per second.

The collection of attribute data through the EIMO equipment, the target blood pressure by SunTech Tango wrist blood pressure measuring instrument records. SunTech Tango wrist is a professional medical blood pressure measurement device, which can be synchronized with software on EIMO devices. The blood pressure prediction experiment was carried out in Inter (R) Xeon (R) E3-1231V3 @3.40G Hz CPU processor. The programming language is python3, and the development environment is Anaconda3-4.4.0. **Figure 4** shows the process of the data acquisition below:

B. SIGNAL ACQUISITION PROCESS

The physical device of the device is used as a sensor platform to record and pre-process the vital signs (PPG, PTT and HR) displayed by the users directly. As shown in **Figure 4**, the device has implemented on PC compression of the signals sent to the application program for display and log recording. The data collected by the device is recorded by the display application program. Simultaneously, the raw data and the low-level measurement data calculated by the device are displayed on the screen, whilst the data logs generated by the application program may also be stored synchronously in the cloud. On the other hand, they are sent directly to service providers electronically for monitoring and large data analysis.

1) PPG Signal Acquisition

PPG data are usually recorded by bedside monitors in hospitals and blood perfusion in tissues is displayed. The signal status is characterized by low frequency components, which are formed near the basic frequency through vascular transmission. The two components of the signal are: the direct current component of the image composed of tissue distance and basic components, and the AC pulse component formed by optical absorption by changing the blood volume of the tissue, This can be detected by the photosensitive element in the PPG sensor device and the signals are received by two reflective PPG sensors. The design of this particular sensor is simple, and the sensing device and light source are all included in the same package.

2) ECG Signal Acquisition

ECG signals are produced by the polarization and depolarization of the myocardium. This potential is generated by the contraction of the main ventricle, with its magnitude and amplitude depending on the position of the electrodes. The whole body loop is then completed by connecting the electrodes with the fingers, and the measurement is achieved by the the sensor device within. The first part of the signal path is received by an ECG sensor, which is Plessey Semiconductor's new potential integrated circuit (EPIC) sensor (PS25201).The power supply of the low noise sensor can help to enable clearer recordings of the ECG signal as the smaller the distortion, the better. The equipment is equipped with DC-DC(Direct current) power supply and low differential output regulator, which can further reduce noise. The sensor can detect and amplify the potential of the surrounding objects, which makes it an ideal choice for detecting tiny electrical interferences in the heart. The device is equipped with two such sensors and is connected with differential amplifiers - the common mode noise between the two sensors is thus reduced and sometimes even eliminated. This results in a potential difference between the left and right hands,

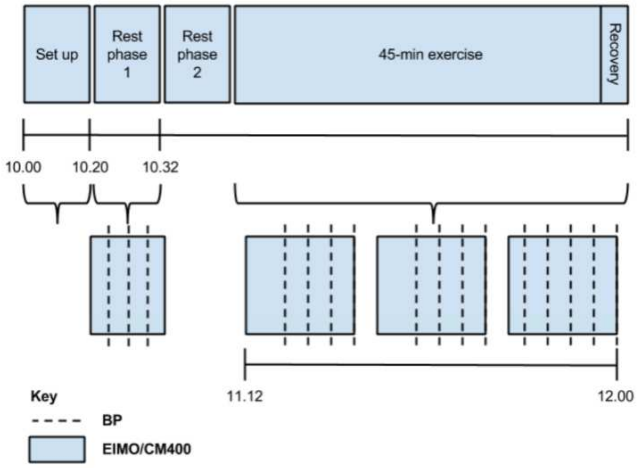


FIGURE 5: Process diagram of movement and rest

which is the same as the 4 lead ECG settings and similar to channel II. The signal is then passed through a band-pass filter before entering the controller ADC.After digitization, the signal is filtered by a 9-order FIR filter: in order to reduce noise and eliminate DC, the limit is 18 Hz to 27 Hz, in order to reduce noise and eliminate DC.

C. DATA ACQUISITION PROCESS

Participants then completed a 45 min's exercise session on a cycle ergometer positioned in an upright seated position similar to the resting posture. Three exercise intensities were used (25, 50, 75W), with participants cycling continuously at each intensity for 15 minutes. After 45 minutes of exercise, the intensity was reduced to 20W for 3 minutes to allow participants time to cool down. Electrocardiography and PPG were measured continuously from the EIMO device during the rest and exercise phases. Twelve BP measures from the SunTech Tango were taken at 2 minutes intervals during the exercise period with one final BP taken after the two minute cooling down period. All BP measures were taken from the right arm. The set-up of the rest and exercise sessions are displayed in **Figure 5**.

D. OPTIMIZATION PARAMETER ANALYSIS OF MODEL

During the experimental process the model is optimized by cross validation for each model (due to the LS (least squares method) possessing no regularization parameter, it is not optimized.) According to the optimized parameters, the prediction model on blood pressure was established, with the prediction analysis being carried out on the test set. The process of finding the optimal parameter for the GBDT of Ps with 1 user is taken as an example. The process is shown in **Figure 6**.

It is observed in **Figure 6** that the abscissa is a sequence of different parameters and the ordinate is scores. For user 1, a peak is reached near the abscissa 105, which corresponds

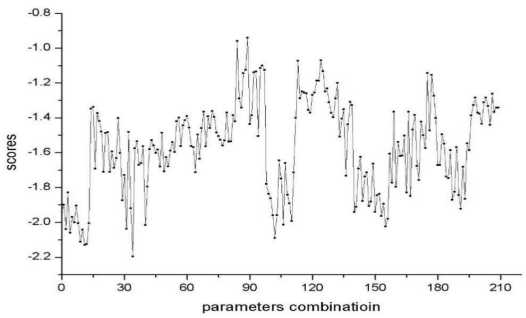


FIGURE 6: Parameter combination

to user 1's optimal parameter combination in table 1. Other people seek parameter processes with user 1.

E. COMPARATIVE ANALYSIS OF EXPERIMENTAL RESULTS

1) Comparative analysis of individual modeling

In order to accurately and effectively test the effect of algorithm, evaluating the performance of the model is based on the statistics of the sample set prediction error in [+5, -5]. The average absolute error evaluation index used to evaluate the classic method outlined in this paper is as follows:

$$accuracy = X/N \tag{8}$$

(1) Accuracy: according to the ANSI/AAMI SP10-1992 of the United States on systolic and diastolic pressure measurement requirements: if the error between the measured ones and the ground truths is less than or equal to 5mmHg, the measurement can be deemed as accurate. In this experiment, we used (-5, +5) mmHg for the accuracy range of measurement errors, i.e. the difference between the predicted blood pressure and the true blood pressure is greater than 5, we treated it as an error; if the predicted blood pressure and the true pressure is less than or equal to 5, we regarded this measurement as correct results.

The acc said the accuracy, X is the number of satisfied blood pressure measurement within the error range and N is the total number of measurement. This experiment will evaluate the accuracy of the prediction using the LS, GBDT, RR (ridge regression), SVR, ElasticNet, KNN and Lasso. The results are shown in **Figure 7** and **Figure 8**, the abscissa respectively are different users.

As shown in **Figure 7** and **Figure 8**, due to differences such as in the way the EIMO devices were held, seating positions and individual posture, the quality of the data gathered is mixed. LS, RR, Lasso, the accuracy of ElasticNet and KNN Ps forecast the worst of 10% to 20%, the best being no more than 47%, The accuracy rate of predicting low blood pressure is between a large floating range of 45% to 70%, due to the differing levels of accuracy with each individual participant's blood. The LS, RR, Lasso and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2902217, IEEE Access

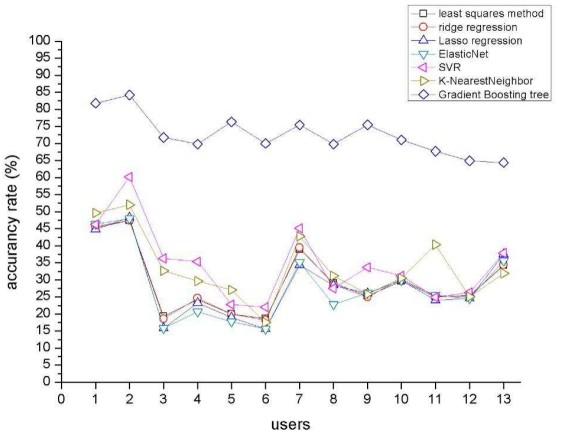


FIGURE 7: The accuracy of Ps is predicted by each algorithm

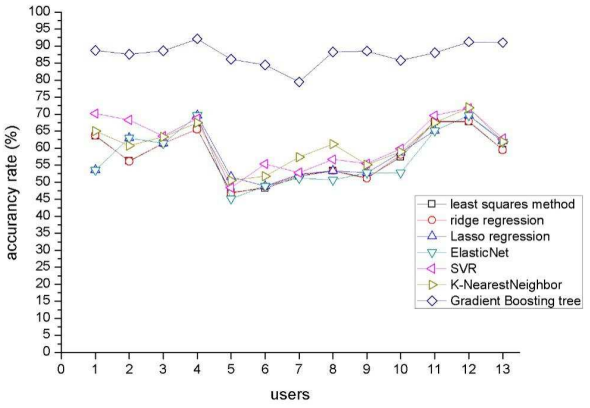


FIGURE 8: The accuracy of Pd is predicted by each algorithm

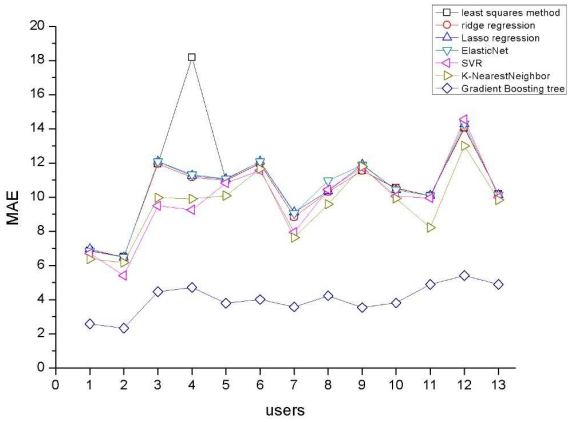


FIGURE 9: The MAE of Ps is predicted by each algorithm

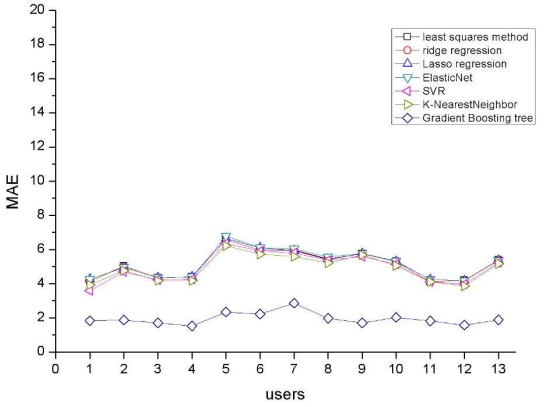


FIGURE 10: The MAE of Pd is predicted by each algorithm

ElasticNet all showcased instability, and participants who displayed poor data results, consequently led to the inaccurate and unsteady patterns in prediction performances. SVR and KNN algorithms promises a slightly better performance rate than the above four algorithms, and in such a case, the accuracy rate is likely to be higher. However, predictions regarding the Ps are the worst in the vicinity of 20%, falling far short of the requirements needed to accurately predict blood pressure. The GBDT displayed the best accuracy rates (with the potential of reaching 90%) in terms of prediction of Ps (stability rates being 65% and above), relatively stable Pd prediction accuracy being more than 80% and the user data quality maintained. Therefore, predicting blood pressure using GBDT is arguably the best method for yielding stable and accurate data.

(2) Mean absolute error(MAE): the The average absolute error is the predictive value of the average absolute value of deviation from the true value. Since the deviation is absolute, the positive and negative offset will not be present. Therefore, the average absolute error is not sensitive to the abnormal value, and reflects the actual situation of the error of the prediction value. The calculation method of mean absolute error is shown in formula (9), in which n represents the number of samples, Y_i representting the actual blood pressure value of the i^{th} sample, with **M** representing the prediction value of the algorithm for the i^{th} sample.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - m| \tag{9}$$

In this paper, the mean absolute error of GBDT algorithm is compared with the LS, RR, SVR, elasticnet, KNN and lasso, The results are shown in **Figure 9** and **Figure 10**.

As shown in **Figure 9** and **Figure 10**, the prediction value of the Pd's MAE gap displayed by the LS, RR, lasso, ElasticNet, SVR and KNN is small, remaining between 3 and 7. The floating range is large, with prediction results varying greatly from user to user. According to the graph,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2902217, IEEE Access

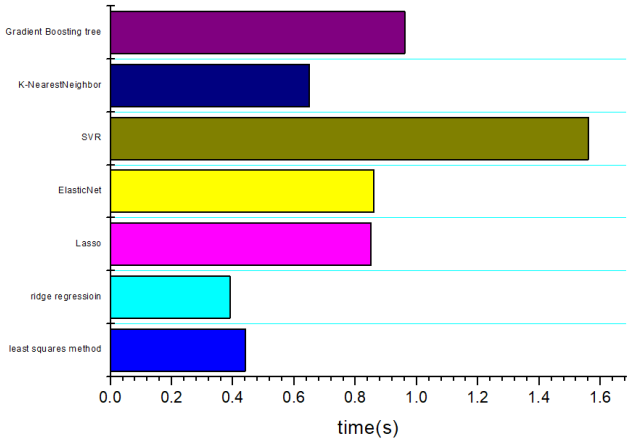


FIGURE 11: Running time of each model

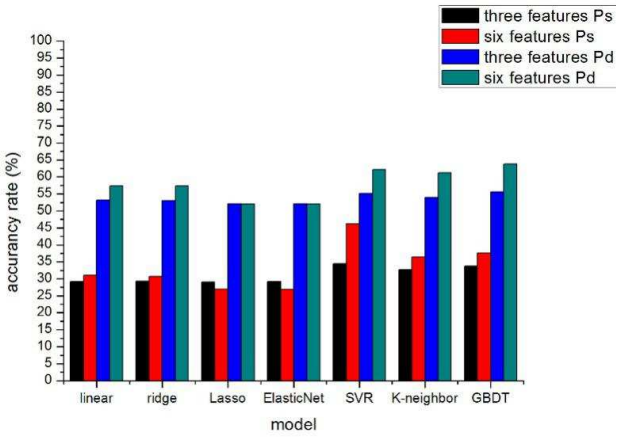


FIGURE 12: the accuracy results for multiple people

using GBDT to predict the Pd's MAE displays evidence of lower accuracy compared to the other six algorithms, with the values remaining below 3. For most users, the MAE value is less than 1.5, displaying a high quality prediction ability. Each algorithm predicts the Ps of MAE, showcasing a big difference. This is due to the change range of Ps is larger than the range of Pd, and the degree of environmental impact is greater. For the LS implemented in the prediction of individual users, the MAE rose to more than 18, and the overall prediction performance is poor. Compared with the LS, the performance of the other algorithms are stable. The overall MAE of SVR and KNN is lower than that of RR, Lasso and ElasticNet, but for individual users, its MAE is still more than 10, with low stability. From the above graphs, a method based on GBDT and one that displays prediction performance rates that surpass that of the other six models may be found, with the worst case scenario involving MAE being no more than 5.5 and the best case scenario involving MAE being 2. Most of the users' MAE is below 4, which means a higher prediction performance rate. This shows that GBDT is the most suitable algorithm for predicting blood pressure.

(3) Running time: Running time refers to the time taken from training the model to outputting the results. In this paper, the running time of each model is the total time trained and tested on the dataset of 13 people. **Figure 11** below described the running time comparison of each algorithm.

In **Figure 11**, the running time of GBDT algorithm is closed to 1s, which is not much better than other algorithms. The prediction time per capita is about 1/13s, which has reached the requirement of realtime prediction. Moreover, the accuracy and MAE of GBDT are the best relative to other models based on these algorithms.

2) Modeling and Analysis for Multiple Individuals

As the EIMO device is not solely intended for personal individual use but is also employed for tracking family health, high accuracy rates in prediction performance ability. Therefore, this experiment attempts to model for multiple people, and observes the prediction performance. The data characteristics regarding the age, weight and height of users 1, 2, 3, 4 and 5 were collected, with the body fat ratio calculated by weight, height and the aforementioned three main data characteristics, which in turn were added to the blood pressure prediction experiment. The formula of body fat ratio (BMI) is as follows:

$$BMI = \frac{height(cm)}{weight(kg)} \tag{10}$$

The information of all the aforementioned users is shown in **Table 1**.

TABLE 1: **Information of the users**

user	height	BMI	age
1	171	1.84	31
2	174	2.03	24
3	164	2.79	23
4	180	2.19	32
5	189	2.55	21
6	165	2.88	27

Experiments are modeled on PPG, HR, PTT and PTT, HR, PPG, age, height, and BMI respectively, with the results shown in **Figure 12** and **Figure 13**.

As shown in **Figure 12** and **Figure 13**, under the condition of using PPT, HR and PTT feature combination. The accuracy of SVR is higher than that of GBDT, but the prediction time of GBDT is less than 0.5s. However, the prediction time of SVR is 1.8s, which is not reflective of prediction performance in real time. Therefore, SVR is unsuitable for modeling the blood pressure rates for multiple individuals. With the new characteristics added to GBDT, tthe accuracy of the Ps is increased by 4.09% and is improved by 8.16%, and the degree of promotion is fairly great. For the MAE index, when using PPT, HR, PTT feature combination, the MAE of each algorithm to predict Ps is higher than 9.5, and the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2902217, IEEE Access

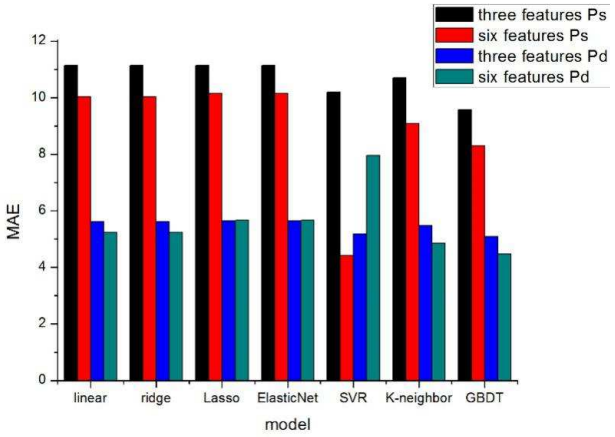


FIGURE 13: the MAE results for multiple people

Pd of MAE close to the value of 5. This shows that GBDT is unable to accurately predict the ability of Ps, but is able to do so regarding measuring Pd. With the addition of new features, the prediction quality of each algorithm's MAE has decreased, indicating the addition of new features increased improve the prediction performance of the algorithm. Among them, GBDT has the lowest MAE in the prediction of Ps and Pd, indicating that GBDT showcased the best performance out of the other six traditional traditional machine learning algorithms in multiple person modeling. With the increase in data, the accuracy of GBDT will see further improvement.

IV. DISCUSSION

The characteristic attribute data and blood pressure data were acquired by EIMO equipment and SunTech Tango sphygmomanometer in respectively real measurement environment. Due to the differing nature of each individual's mode of operation and power measurement of the device battery amongst other objective factors, the data quality varies from participant to participant. Regarding the test subjects who displayed high quality data, the algorithm performed better in terms of prediction ability. In this paper, the method of cross validation running in the background for the training data, whilst the optimal combination of the parameters are used to carry out the procedures of prediction. The SVR algorithm's actual prediction time for a single individual is less than 1s, whilst for multiple individuals, the resulting time is more than 1.8s. For other algorithms, not only for all but single, the prediction time is less than 0.5s, showcasing high effectiveness. On a final note, the data yielded from the experiments is based on healthy individuals aged between 20 and 30 years. Consequently, this calls for a more varied pool of participants regarding age range for future research opportunities. Algorithm performance, especially of that concerning measuring the blood pressure rates of multiple participants, also showcases improvement when additional, new features regarding age, height and BMI are included. For future studies and especially greater algorithm accuracy, the aim is to also include data related on alcohol intake, family medical history and other relevant factors related to the algorithm training.

V. CONCLUSION

For the findings outlined in this paper, the EIMO device is used to collect the physiological data from the human body, alongside technological devices such as the Wrist Blood Pressure Monitor and Suntech,which focused on collecting blood pressure values. GBDT algorithm was them employed to provide analysis and modeling for the data, predicting blood pressure rates in the process.In terms of evaluating the performance quality of the GBDT algorithm, this was tested against the traditional LS, RR, SVR, ElasticNet, KNN and Lasso. Based on the overall accuracy rate and average absolute error evaluation, results showcased that the GBDT algorithm yielded the best performance - the training time is less than 0.5s, making it suitable for making predictions in real time scenarios. In addition, the portability of the EIMO equipment and its ability to retain physiological data over an extended period time also makes this method suitable for accurately measuring blood pressure regardless of the environmental and time constraints. Regarding predictions based on multiple individuals, this paper combined the characteristics of age, body fat ratio and height and this greatly enhanced the performance of the algorithm in terms of accuracy rate. We therefore conclude that adding more feature attributes and individualised data will further improve the algorithm's efficiency.

REFERENCES

[1] Gesche H, Grosskurth D, KÃijchler G, et al. Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method[J]. European Journal of Applied Physiology, 2012, 112(1):309-315.

[2] Hsieh Y Y, Wu C D, Lu S S, et al. A linear regression model with dynamic pulse transit time features for noninvasive blood pressure prediction[C]// IEEE Biomedical Circuits and Systems Conference. IEEE, 2016:604-607.

[3] Elamvazuthi I, Aymen O M, Salih Y, et al. An intelligent control of Blood Pressure system using PID and Neural Network[C]// Industrial Electronics and Applications. IEEE, 2013:1049-1053.

[4] Li X, Wu S, Wang L. Blood Pressure Prediction via Recurrent Models with Contextual Layer[C]// International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017:685-693.

[5] Abbasi R, Moradi M H, Molaeezadeh S F. Long-term prediction of blood pressure time series using multiple fuzzy functions[C]// Biomedical Engineering. IEEE, 2015:124-127.

[6] Peng R C, Yan W R, Zhang N L, et al. Cuffless and continuous blood pressure estimation from the heart sound signals[J]. Sensors, 2015, 15(9): 23653-23666.

[7] Kurylyak Y, Lamonaca F, Grimaldi D. A Neural Network-based method for continuous blood pressure estimation from a PPG signal[C]//Instrumentation and Measurement Technology Conference (I2MTC), 2013 IEEE International. IEEE, 2013: 280-283.

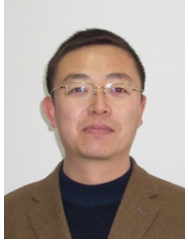
[8] Sharma M, Barbosa K, Ho V, et al. Cuff-Less and Continuous Blood Pressure Monitoring: A Methodological Review[J]. 2017, 5(2):21.

[9] Munnoch R, Jiang P. A personal medical device for multi-sensor, remote vital signs collection in the elderly[C]// Science and Information Conference. IEEE, 2015:1122-1131.

[10] Piper M A, Evans C V, Burda B U, et al. Diagnostic and predictive accuracy of blood pressure screening methods with consideration of re-

screening intervals: a systematic review for the U.S. Preventive Services Task Force.[J]. Annals of Internal Medicine, 2015, 162(3):192-204.
[11] Allen J. Photoplethysmography and its application in clinical physiological measurement[J]. Physiological measurement, 2007, 28(3): R1.
[12] Teng X F, Zhang Y T. Continuous and noninvasive estimation of arterial blood pressure using a photoplethysmographic approach[C]// Engineering in Medicine and Biology Society, 2003. Proceedings of the, International Conference of the IEEE. IEEE, 2004:3153-3156 Vol.4.
[13] Aoyagi T, Fuse M, Kobayashi N, et al. Multiwavelength pulse oximetry: theory for the future[J]. Anesthesia & Analgesia, 2007, 105(6 Suppl):S53.
[14] Poon C C, Zhang Y T. Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time[C]// Engineering in Medicine and Biology Society, 2005. Ieee-Embs 2005. International Conference of the. IEEE, 2005:5877-5880.
[15] Jadooei A, Zaderykhin O, Shulgin V I. Adaptive algorithm for continuous monitoring of blood pressure using a pulse transit time[C]// Electronics and Nanotechnology. IEEE, 2013:297-301.
[16] Ye S Y, Kim G R, Jung D K, et al. Estimation of systolic and diastolic pressure using the pulse transit time[J]. World Academy of Science Engineering & Technology, 2010(67):726.
[17] Chen W, Kobayashi T, Ichikawa S, et al. Continuous estimation of systolic blood pressure using the pulse arrival time and intermittent calibration[J]. Medical & Biological Engineering & Computing, 2000, 38(5):569.
[18] Obrist P A, Light K C, Mccubbin J A, et al. Pulse transit time: relationship to blood pressure and myocardial performance[J]. Psychophysiology, 1979, 16(3):292.
[19] Wong M Y M, Zhang Y T. The relationship between pulse transit time and systolic blood pressure on individual subjects after exercises[C]//Distributed Diagnosis and Home Healthcare, 2006. D2H2. 1st Transdisciplinary Conference on. IEEE, 2006: 37-38.
[20] Zhang B , Wei Z , Ren J , et al. An Empirical study on Predicting Blood Pressure using Classification and Regression Trees[J]. IEEE Access, 2018, 6: 21758-21768.
[21] Hastie T, Tibshirani R, Friedman J H. Elements of Statistical Learning[J]. Springer, 2001, 45(3):267-268.
[22] Zurek P, Krejcar O, Penhaker M, et al. Continuous Noninvasive Blood Pressure Measurement by Near Infra Red CCD Camera and Pulse Transmit Time Systems[C]// Second International Conference on Computer Engineering and Applications. IEEE Computer Society, 2010:449-453.
[23] Nagelkerke N J D. A Note on a General Definition of the Coefficient of Determination[J]. Biometrika, 1991, 78(3):691-692.

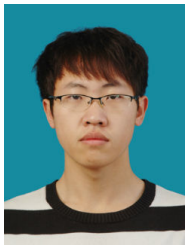
BING ZHANG is currently a Lecturer and holding a post-doctoral position at the School of Information Science and Engineering, Yanshan University, China. He received his bachelor's degree from the College of Computer and Information Technology, Three Gorges University, China, in 2012, and the Ph.D. degree from the School of Information Science and Engineering, Yanshan University, China, in 2018. He has ever been with the Norwegian University of Science and Technology as a Visiting Scholar. His research interests include data mining, machine learning, and software security.



JIADONG REN is a professor in the school of Information Science and Engineering, Yanshan University, China. His research interests include data mining, temporal data modeling and software security. His research has been supported by National Natural Science Foundation of China and Science Foundation of Hebei Province. He is a senior member of the Chinese Computer Society, member of IEEE SMC Society and ACM.

YONGQIANG CHENG is a Senior Lecturer in the department of computer science and technology, University of Hull, UK. Before this, he was a postdoctoral scientific researcher in the Future Ubiquitous Networking Laboratory in the School of Engineering and Informatics, University of Bradford since 2010. He received both his bachelor and master degrees in control theory and control engineering from Tongji University, Shanghai,China, in 2001 and 2004, respectively. He obtained his Ph.D. degree in 2010 from the School of Engineering, Design and Technology, University of Bradford, UK. His research interest includes artificial intelligence, machine learning, smart systems and digital health, control theory and applications, embedded systems etc.

BING WANG has been the leading researcher in the areas of database theories and mark-up languages semantics and design. He is currently the university Lecturer at the department of computer science and technology, University of Hull, UK. He received his PhD from the University of York, England and did his Post-doctoral research at the Engineering Department, University of Cambridge, where he designed and developed a multi-task multimedia interface which is now widely used for The British Petroleum Company plc, UK. His currently research interests include semantic data modelling, hypermedia systems and medical related technologies.



ZHIYAO WEI is currently a graduate student and received his bachelor's degree in 2016 in the school of Information Science and Engineering, Yanshan University, China. His research interests are machine learning and data mining. He's now worked on a project on medical data.